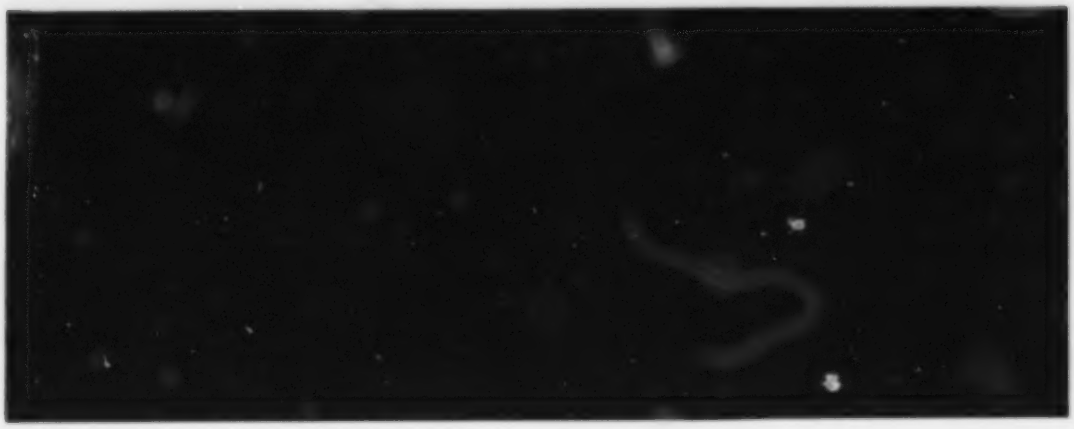
Statistics Canada | Statistique Canada



Methodology Branch

Census & Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes de recensement
et d'enquêtes ménages

Canada

E-624E-C2

SOME METHODOLOGICAL ASPECTS OF THE 1986 CENSUS OF CANADA

Paper presented at the Statistics Canada - US Bureau of the Census Interchange

D. Royce

Working Paper No. CHSMD- 85 - 037 - E

1. INTRODUCTION

The 1986 Census is now less than 19 months away, and preparations at Statistics Canada are well underway. Following a two-year process of consultations and negotiations between Statistics Canada and federal departments, provincial governments and other major user groups, the content of the 1986 Census was approved by the Cabinet in late June of this year. Key managers in the Census Management Team were appointed beginning in mid 1983, and planning and development activities are now shifting into high gear.

This report highlights some of the more interesting methodological issues that have arisen or which are foreseen for 1986. The paper emphasizes issues related to data collection, as this is the area where most of the work has been concentrated in the past 12 months. Coverage and data quality aspects are not covered, however. These will be the subject of the presentation by R. Burgess later in this session.

The context within which initial planning for the 1986 Census took place was that the 1981 Census was considered to be an unprecedented success. The data were collected and processed within budget, and the results were released a year ahead of the 1971 Census schedule. Consequently, the basic planning premise was that the 1981 methodology and procedures should be repeated wherever possible, with changes being made only to correct known deficiencies in data quality or where substantial savings in cost or time could be realized. Before describing some of the new issues for 1986, a general overview of the 1981 Census methodology will be given.

2. THE METHODOLOGY OF THE 1981 CENSUS

2.1 Collection

The Canadian Census is actually two censuses; a Census of Population and Housing and a Census of Agriculture. The data collection phases of the two censuses are combined, although the questionnaires are later split and processed separately. By conducting them simultaneously, there are considerable cost savings compared to conducting a separate Census of Agriculture. In what follows, we restrict ourselves to a description of the Census of Population and Housing.

The first step in the Census involves the geographical delineation of the entire country into over 40,000 Enumeration Areas (EAs). For each EA a map is prepared and a Census Representative (CR) is hired and specially trained to carry out the collection activities. Three different collection methodologies are used:

(i) Dropoff - Mailback

This method is used in urban centres, which constitute about 63% of the population. The CR lists all dwellings and drops off a census questionnaire at occupied dwellings. This is is sometimes referred to as the "list-leave" technique. The householder is asked to complete the questionnaire on Census Day and return it by mail.

(ii) Dropoff - Pickup

This method is used in small urban and rural areas. The dropoff phase is exactly the same as in mailback areas. However, the CR returns to each household to pick up the completed form. The method applies to approximately 35% of the population.

(iii) Canvasser

This method is used for remote areas or where self-enumeration would be difficult, for example due to illiteracy or language problems. It is used for about 2% of the population.

In mailback areas, once the questionnaires are received back, they are subjected to edit and follow-up by the same CR who did the dropoff. The CR edits each questionnaire and follows-up in an attempt to obtain complete response where required. Certain questions are specified as requiring mandatory follow-up, but follow-up may also be initiated if too many questions on one form fail edit. Follow-up is conducted by telephone whenever possible, otherwise field follow-up is used. Complete nonresponse is also followed-up in the field.

In pick-up areas, the editing of questionnaires is conducted at the time

the CR picks up the questionnaire and follow-up is conducted on the doorstep.

Once the CR has completed his/her assignment, it is checked by the CR's supervisor, the Census Commissioner (CC). Once the Census Commissioner is satisfied, the EA is passed to the Field Quality Control Technician, who conducts a number of checks for quality. Checks of questionnaire editing are done using sample verification. If the EA is rejected it is returned to the CC and the CR for further correction. Once the EA has passed Quality Control, it is passed on to Regional Office Processing.

2.2 Sampling

At this point it is convenient to describe how sampling was used during data collection.

For private households, there were two questionnaires, a short (Form 2A) and a long (Form 2B). The short form contained basic demographic questions, mother tongue, coverage-related questions, tenure and structural type of dwelling. The long form contained additional questions on housing, place of birth, mobility, ethnic origin, labour force activity, education and income. One in five households was to receive the long form.

To implement the sampling, every fifth line in the CR's Visitation Record was shaded. The Census Commissioner was given a random number indicating whether none, one, two, three or four lines at the beginning of the Visitation Record were to be crossed out. The CR was instructed to list dwellings in the EA by starting at a pre-specified location and proceeding in a specific direction. The CR then dropped off a long form at every household on a shaded line.

2.3 Coverage Improvement Methods

Complete coverage is the primary objective of the census, and so there were a number of steps taken to enhance it in 1981. Map checks were carried out prior to enumeration, an extensive publicity campaign was mounted, there were checks of the CR's list of dwellings against Post Office records, dwelling counts were checked against expected counts for large discrepancies, and questions were included on the questionnaire intended to identify persons

likely to be missed. Because of their importance, two other procedures are described in further detail below.

The Canadian Census is a de jure census, that is, each person is to be enumerated at his/her usual place of residence. It is known that a large number of persons will not be at their usual place of residence on Census Day and may be missed. Each of these "temporary residents", wherever they are on Census Day, whether staying in a hotel, visiting relatives, or wherever, was required to complete a special questionnaire (Form 3) and to indicate his usual place of residence. During processing, a sample of these forms was selected and matched back to their reported usual place of residence to see if they were enumerated there. From this sample, the number of temporary residents who were missed was estimated, and this estimate was used to adjust the final population count.

Another known source of coverage error is the misclassication of occupied dwellings as unoccupied. A special study, called the Vacancy Check, was conducted to estimate the extent of this error. It involved selecting a sample of dwellings originally classified as unoccupied and following them up using specially trained interviewers shortly after the census. The estimates of undercoverage from this study were also used to adjust the final population count.

2.4 Regional Office Processing

Completed questionnaires for each EA were forwarded to one of five Regional Processing Offices where a series of six operations were carried out:

Operation 1 - Receipt and Registration

This consisted of receipt of EA boxes and their registration in a master control log which tracked their progress through subsequent operations.

Operations 2 - Reverse Record Check

Only a sample of EA boxes was used in this step. The Reverse Record Check is our major vehicle for estimating undercoverage. The operation

involved searching through 1981 documents for a sample of persons selected from 1976 Census records.

Operation 3 - Document Preparation

This operation involved a check of the consistency of information on questionnaires with the corresponding Visitation Record, to ensure that certain basic counts agreed. In addition, documents were groomed for key-entry and written responses for household status, language, place of birth, ethnicity and religion were coded. Statistical quality control was done on all editing and coding.

Operation 4 - Additions and Sequencing

This operation consisted of a reconciliation of the totals from questionnaires and the Visitation Record for the EA, and the establishment of total documents, households and persons as control totals for subsequent operations. Documents were also sequenced by type to assist in further coding and key-entry.

Operation 5 - Geographic and Economic Coding

Write-in answers to the Mobility Status, Industry, Place of Work, Occupation and Class of Worker questions were coded in this operation. Again, statistical quality control similar to that used in Operation 3 was applied.

Operation 6 - Batching and Labelling

Documents were batched by type for the purpose of keying and were then shipped to Direct Data Entry.

2.5 Direct Data Entry

In the 1971 and 1976 censuses, the FOSDIC system developed by the U.S. Bureau of the Census was the method used for data entry. However, for the 1981 Census the decision was made to use existing key-entry facilities at Revenue Canada (our counterpart of your IRS) for our data entry. The time period for our data entry operation coincided very well with Revenue Canada's low period in the income tax processing cycle, and we were able to use Revenue Canada's very experienced personnel. The keying was done in the same regional

offices where the manual processing was carried out, and this had some advantages of better communication. Although the keying and verification themselves were done regionally, the terminals were all connected to a central computer in Ottawa where all editing and data storage was done.

2.6 Head Office Processing

The three main inputs to the Head Office Processing operation were the Visitation Records, the data tapes from Revenue Canada, and the questionnaires for each EA. Receipt of this material was registered on a computer terminal and an automated control system known as the Process Control Master File recorded the status of each EA through Head Office Processing and ensured that no steps were overlooked.

Once registration was complete, Visitation Records were sent to our Geography Division to make sure each EA boundary was respected. Corrections were made on both the documents and the database.

Next, the data tapes were subjected to a set of structural edits for document and household types, the reconciliation of document and person counts, and a few other checks.

The final operation in Head Office Processing involved the special processing of temporary resident documents (Form 3) and documents for merchant and naval seamen, personnel in the armed forces, and personnel in Canadian Embassies outside the country.

2.7 Edit and Imputation

Following Head Office Processing, the data were loaded onto a RAPID data base and processed through Edit and Imputation. Two separate computer systems were used in 1981. The CAN-EDIT system, based on the methodology first developed by Fellegi and Holt (1976), was used for processing all 2A data, except for mother tongue, and for several of the 2B variables. This system uses edit rules specified as a set of invalid conditions or conflict rules. The system analyses the edit rules, identifies conflicts and redundancies, and generates all implied edits. The imputation actions are then driven directly by the edit rules using the principle of minimizing the number of changes required to

satisfy the edits. The actual imputation is done by means of a hot deck approach.

A major limitation of CAN-EDIT, however, is that it cannot handle variables with long code lists due to the large number of edit rules required. It also does not have the flexibility to handle variables where there are known error patterns requiring very specific correction actions.

The second system, referred to as SPIDER ("System for Processing Instructions from Directly Entered Requirements") was used for edit and imputation of all other variables. The "directly entered requirements" refers to the set of subject-matter edits, specified in the form of decision tables. The SPIDER system translated these decision tables into PL/1 and packaged them into a processor which performed the function specified in the tables. Corrections to "bad" records identified by SPIDER could be made either according to prespecified rules (deterministic correction) or by searching clean records for a suitable donor record. The matching requirements for finding a suitable donor were partly determined by the edit rules which the bad record failed and partly by matching constraints specified by the subject matter specialist. Thus, SPIDER permitted a higher degree of "hands-on" manipulation of the data than did CAN-EDIT.

2.8 Weighting

An additional step for the 2B (sample) variables was the weighting of the data. This was accomplished using the Raking Ratio Estimation procedure. The data were first sub-divided into approximately 5000 geographic Weighting Areas (WAs) with an average population of 5000 persons in each area. For each WA, a two-dimensional weighting matrix was formed, with the rows and columns of the matrix defined by selected 2A variables or combinations of variables. The sample counts and population counts were entered into the matrix, collapsing of rows and/or columns was done if necessary, and the weighted sample counts on the margins of the matrix were iteratively fit to the corresponding population counts until the weights converged or until a specified number of iterations was exceeded.

Before being loaded onto the retrieval database (also a RAPID database),

the adjustments to the population counts mentioned earlier were implemented. Records were selected at random according to the undercoverage estimates from the Temporary Residents Study and the Vacancy Check, and the weights of the selected records were increased.

We now turn to some of the new developments for 1986.

3. ABORIGINAL PEOPLES DATA ISSUES

The new Canadian Constitution of 1982 contains specific provisions regarding the rights of aboriginal people, with further negotiations scheduled to take place over the next several years. A large and increasing number of government programs are now directed to native peoples.

In the 1981 Census, a special User Summary Tape and microfiche on the native population were produced in response to user demand. For the 1986 Census, there was a strong demand for a clearer identification of aboriginal persons, as well as a need to provide better data on Indian Reserves.

3.1 Question 7

The first of these needs was addressed by the addition of a new question on the short (100%) form directed at aboriginal people and allowing them to specify their native group. The question reads as follows:

7. Do you consider yourself an aboriginal person or a native Indian of North America, that is, Inuit, North American Indian or Métis?

__ No, I do not consider myself Inuit, North American Indian or Métis
__ Yes, Inuit
__ Yes, status or registered Indian
__ Yes, non-status Indian
__ Yes, Métis

From previous censuses, it was known that only about 2% of the Canadian population could be expected to provide a positive response to such a question. The possibility was raised that there might be a relatively high non-

response rate on the part of the other 98% of the population, since they might leave the question blank if the condition did not apply to them, rather than check the "No" box as requested. This hypothesis was supported by the U.S. experience with the Hispanic origin question in the 1980 Census, which had a 15.5% nonresponse rate prior to follow-up and a 4.2% rate after follow-up, the highest of all 100% questions.

The issue of whether to conduct follow-up when this question alone was left blank thus arose. To conduct follow-up would cost money, but how much was unknown since no estimate of the expected nonresponse rate to the question was available. However if the question were not followed up, the problem would be left to imputation to resolve, with the possibility of unknown imputation biases.

As a result, a small-scale test of the proposed 1986 short form questionnaire was conducted in September 1984, using a subsample from the monthly Labour Force Survey. Questionnaires were mailed out to approximately 1700 households, followed by a thank you/follow-up letter 2½ weeks later. The final mail return rate was 45.2%. Each returned questionnaire was edited following the edit steps to be used by enumerators in 1986, and was then classed as either passing edit, failing edit only for Question 7, or failing edit for some other reasons. In addition, nonresponse rates were tabulated for each question on the short form. The test results indicated that the nonresponse rate to Question 7 was indeed much higher than for any other individual question (Table 1). An estimated 7.3% of short form questionnaires had nonresponse only to this question.

Table 1: NON-RESPONSE RATES FOR INDIVIDUAL QUESTIONS, 1984 CENSUS TEST

QUESTION	NON-RESPONSE[1] RATE (%)
Name	0.9
Relationship to Person 1	1.2
Date of Birth	1.0
Sex	1.7
Marital Status	3.1
Mother Tongue	4.7
Native Peoples	16.5

[1] Includes multiple responses to relationship, sex, marital status, mother tongue and Yes/No combinations for Native Peoples question

The results of this test will allow a more precise estimate of the costs of follow-up to be made, and indicate that some further work is necessary if the nonresponse rate to this question is to be lowered.

3.2 Data for Indian Reserves

The possibility of some form of self-government for aboriginal peoples, not unlike that for municipal governments, has made better quality data for Indian Reserves a priority as well. Better quality data will be required in terms of better coverage, better quality response to the questionnaire, and more reliable long form data at the level of the individual Indian Reserve. Since the majority of Indian Reserves have less than 1000 population, with many under 500, one suggestion that has been made in regard to the latter is to do 100% sampling of the long form on Indian Reserves.

An analysis of imputation rates for Indian Reserves in 1981 indicated that, in general, reserves tended to have higher nonresponse rates than did the non-reserve population, even for the same collection methodology (Table 2). However, the high nonresponse rates were concentrated in relatively few EAs. The findings were similar to those of studies conducted after the 1971 and 1976 censuses. Nonresponse was especially high for variables with conceptual difficulties in relation to the native population, such as shelter costs, place of work, and income (Table 3).

Table 2: IMPUTATION RATES[1] FOR INDIAN RESERVES AND NON-INDIAN RESERVE AREAS, BY COLLECTION METHODOLOGY, IN PERCENT

	COLLECTION METHODOLOGY			
Area Type	Mailback	Canvasser	Pickup	Overall
Indian	5.5	10.6	6.7	9.6
Non-Indian	4.2	6.5	3.4	4.0

[1] after removal of total refusal Indian Reserves

Table 3: IMPUTATION RATES[1] FOR SELECTED VARIABLES, BY INDIAN RESERVES AND NON-INDIAN RESERVE AREAS, IN PERCENT

	AREA TYPE	
Variable	Indian	Non-Indian
Place of Birth, Citizenship	6.0	1.9
Income	14.6	8.4
Mobility	5.9	1.2
Official Language	10.2	2.8
Place of work	22.1	11.6
Religion	3.6	1.0
Shelter Costs	53.3	19.0
Ethnic Origin	1.1	1.5
Home Language	7.1	2.3

[1] after removal of total refusal Indian Reserves

Problems were also encountered on Indian Reserves in 1981 related to sampling . In several cases, large deviations from the specified 1 in 5 sampling ratio were observed. In some cases, this occured due to a large number of total refusals to the long form. In such cases, the document type was artificially changed to a short form, thus leading to a less than 1 in 5 ratio. In other cases, reserves originally designated for self-enumeration and sampling were changed to canvasser and 100% long form but the changes were not entered into the processing system, thus leading to a greater than 1 in 5 ratio. However, since Indian Reserves form only a small part of a weighting area where the overall sampling ratio is close to 1 in 5, weights close to 5 were assigned. When sample data were then tabulated for Indian Reserves using these weights, large inconsistencies between population and weighted sample counts were observed.

As well as deviations from the sampling ratio, there may also be bias in the way the sample is selected. A non-scientific perusal of the documents for some reserves has suggested that the interviewer tended to drop the long form at the homes of better educated and employed respondents. Further work would be needed to confirm this suspicion, however.

As mentioned previously, one suggestion to improve the situation is to switch to 100% long form on Indian Reserves. Although this would certainly reduce the sampling variance, nonresponse bias and response errors could still remain as serious problems. Furthermore, the legal basis for such a decision is not clear. Other proposals include the use of the canvasser method as the standard collection methodology to circumvent the high rates of illiteracy encountered on reserves. It has also been suggested that the CR concentrate only on the Census of Population, with Agriculture handled separately, and that the CR be given more extensive training on the questions themselves and their interpretation in special circumstances. Indian Reserve data profiles may also be developed to assist field personnel in convincing Indian Band Managers of the value of the census at a local level. Changes to processing which would see Indian Reserves treated separately during imputation and weighting are also under consideration.

Improved data on the aboriginal population is a challenging assignment.

Previous data collection efforts by the Census and by other Statistics Canada surveys have not been completely successful. It will be interesting to see how well we do.

4. EXTENSION OF BILINGUAL SERVICES

As an agency of the federal government, Statistics Canada is committed to offering services to the public in the official language (English or French) of their choice. An important application of this policy is offering respondents the option of completing the census questionnaire in the language of their choice.

In 1981, as in previous censuses, this policy was implemented while using separate English and French questionnaires. If contact with the householder was made at dropoff, a questionnaire in his/her preferred official language was left. If no contact was made, the procedure varied depending on whether or not the area was designated as bilingual. Bilingual areas were defined as those area where 10% or more of both official language group were present, and accounted for about 13% of the population. In such areas, a so-called double dropoff procedure was used which involved leaving behind both an English and a French questionnaire. In all other areas, a single unilingual questionnaire was left behind which was either in the householder's preferred language (if that could be determined), or in the language of the majority group in the area. In addition, a toll-free Telephone Assistance Service was available to all residents which could be used to request a questionnaire in the other official language.

These procedures were judged to be very successful. Only 0.1 percent of households actually used the telephone service to request a second questionnaire, and only 17 complaints were lodged with the Commissioner of Official Languages as opposed to 58 in 1971.

For 1986, it was decided to extend the provision of bilingual services still further through means of:

a) use of a bilingual short form (2A) instead of separate English and French forms.

b) the extension of double dropoff procedures for the long form (2B) and the Census of Agriculture questionnaire from designated bilingual areas to all areas of Canada. For operational reasons mainly to do with postal costs, it is not feasible to create a bilingual long form or agriculture questionnaire.

From an operational perspective, the major concern is public reaction to such a change. Negative reactions could occur because of:

(i) the additional cost of such a change (estimated at $1.1 million) when the 1981 procedures were seen as effective,

(ii) environmental concerns over the extra paper required,

(iii) confusion of respondents not used to dealing with bilingual forms,

(iv) a higher perceived respondent burden caused by respondents receiving double the number of pages actually needed,

(v) the feeling that the federal government is being over-zealous in promoting bilingualism where it is not needed.

Such negative reactions could translate into lower mail response rates, increased follow-up, lower CR morale, and higher costs. On the other hand, the reaction of the public could be very favourable if bilingualism is seen as popular at the time.

It is, however, virtually impossible to conduct any methodological test which would accurately predict the public's reaction, since it is likely to depend heavily on factors such as media reaction and the political and economic climate at the time of the 1986 census. Probably the best that can be done is to measure the behaviour of individual respondents.

Some data of this nature are available from an experiment conducted in 1976. A sample of households was selected in areas where it was believed that a bilingual questionnaire would cause problems. A random half of the sample was mailed a bilingual short form questionnaire, while the other half was mailed a questionnaire in the majority language of the area. Although the

response rates did not vary significantly between the two versions of the questionnaire, this finding was limited by the fact that the overall mail return was substantially lower than that obtained in a full census.

In addition to concerns about public reaction, there are other concerns related to coverage, data quality and operational procedures, for example:

(i) Some bilingual respondents may complete both the English and French questionnaires, leading to possible double counting unless checks are made.

(ii) In a similar vein, some households may use both sides of a 2A or both 2B forms due to more than six persons in the household or because persons with different language preferences reside in the household.

(iii) There were cases in the September 1984 test mentioned earlier where the respondent tore or cut the 2A in half and returned only the half that related to their language preference. Some procedure will be required to deal with this.

During the course of the next year, we hope to develop and put in place mechanisms to measure the effect of this change in collection methodology.

5. THE POST CENSAL DISABILITY SURVEY

In response to the International Year of the Disabled, the Government of Canada established a Special Parliamentary Committee on the Disabled and Handicapped. One of the recommendations of the Committee's Obstacles report (1981) was that Statistics Canada develop an improved data base on the disabled and handicapped population. Statistics Canada has met the short-term needs for such data through Labour Force Survey supplements, but the sample sizes are too small to satisfy detailed data needs.

The 1986 Census of Population long form will contain a two-part screening question to identify the disabled population, with a view to using this information to conduct a post-censal survey of the disabled using the Census as a

sampling frame. The exact questions read as follows:

27.(a) Are you limited in the kind or amount of activity that you can do because of a long-term physical condition, mental condition or health problem?

(i) At home?

__ No, I am not limited
__ Yes, I am limited

(ii) At school or at work?

__ No, I am not limited
__ Yes, I am limited
__ Not applicable

(iii) In other activities, e.g., transportation to or from work, leisure time activities?

__ No, I am not limited
__ Yes, I am limited

(b) Do you have any long-term disabilities or handicaps?

__ No
__ Yes

One requirement for a post-censal survey based on these questions is that a sample of both screened-in and screened-out persons be followed up. The Census screening questions themselves are not sufficiently precise to define the disabled population. The criteria most widely recognized and the one adopted by Statistics Canada is the Activities of Daily Living (ADL) scale, a set of 17 questions designed to determine how well a person copes with everyday chores such as carrying a 12 pound bag of groceries 30 feet, cutting their own toenails, etc. The census questions will thus only serve as a stratification variable to improve the efficiency of the post-censal survey's sample

design - a classic example of double sampling for stratification (Cochran, 1977). Persons screened out by the Census questions must be followed up as well as those screened in, but in smaller proportions.

Work began on the post-censal Disability Survey with a two-day workshop held in the spring of 1984. Subsequently, a working level task force was established to look into the feasibility and implications of integrating the Disability Survey into the Census, as opposed to conducting it as a separate survey several weeks or months after the Census. The task force considered a range of integration options, the major ones being:

(i) A fully integrated methodology, where CRs in designated EAs would conduct the interview with screened-in respondents. In mailback areas, the CR would attempt to conduct the Disability Survey by telephone, but if this were not successful a personal interview in the respondent's home would be arranged. In pickup areas, the Disability Survey would be conducted immediately following the doorstep edit and follow-up for the Population and Agriculture questionnaires. In both types of areas interviews would be done by proxy for children, but non-proxy interviewing would be attempted for adults and seniors. For canvasser areas, proxy response for all age groups would be acceptable.

(ii) A partially integrated approach, where CRs in designated EAs would merely prepare a list of all respondents screened in. The actual interviewing would be done by a team of specially trained interviewers. In canvasser areas, the Disability Survey would still be conducted by the CR.

(iii) A non-integrated approach, using a separate field staff and with sample selection occuring after Quality Control of the EAs but prior to data entry. The sample selection would be done by a clerical staff in a separate operation during Regional Office Processing.

With all three options, it was clear that the most efficient way of conducting interviews with the sample of screened out respondents would be by telephone from the Regional Offices. Even though the interview in the majori-

ty of cases would be confined to the 17 questions of the ADL, for a CR to conduct even such short interviews with the hundreds of screened out persons in his/her EA would impose too high a workload. Having the CR subsample screened out respondents within the EA was judged to be too risky, since the CR is already responsible for implementing the sampling of every fifth dwelling for the long form.

With both the fully integrated and partially integrated approaches, the inability of the CR to subsample within EAs also mean that the smallest possible cluster for interviewing the screened in population would be the EA. For the portion of the sample selected in the Regional Offices, i.e., the screened out population under all three options plus the screened in population under the non-integrated option, a two stage design using EAs at the first stage and households or respondents at the second stage would be possible.

With all three options, EAs would be stratified geographically prior to selection and the sampling rate in each stratum would be adjusted to provide sufficiently reliable estimates for a variety of areas such as Census Metropolitan Areas, Census Agglomeration and Census Divisions.

As well as the sample design possibilities just described, the three options were compared on the basis of a number of other factors, including:

1) timeliness of the frame of screened in respondents
2) degree of control over the size of the sample
3) control over creation of the frame by CRs
4) ability to collect data for difficult to enumerate areas
5) experience of interviewers needed for a potentially sensitive and difficult survey
6) privacy concerns when respondent and interviewer know each other
7) respondent burden - concentrated or spread over time
8) timeliness of the final results
9) impact on the Census of Agriculture
10) public relations aspects of using the Census as a sample frame
11) costs

The recommendation made to senior management in early October was that the partially integrated approach was to be preferred. It appears to combine the best balance of the major factors of timeliness, control over the sample size, experience of interviewers, and costs. The next step in the process will now be to prepare a submission to the federal cabinet requesting funding for the Disability Survey.

As a final note, an interesting side issue concerns the appropriate use of the census questions as data items in their own right. Although the census questions are primarily intended as a screening device and are known to be biased when compared to the more precise ADL scale, the census questions will provide more geographically disaggregated data than will the Disability Survey. By following up a sample of screened out respondents, the Disability Survey will be able to produce an estimate of this bias for large areas such as provinces and Census Metropolitan Areas. If this bias can also be related to other characteristics of respondents, it may be possible to use small area data techniques to generate improved estimates for small areas. This will be an interesting area of research in the next few years.

6. AUTOMATED CODING

The final topic I would like to cover in this report is that of automated coding. While strictly speaking this is an aspect of our 1991 planning, we have spent a considerable amount of time on this in the past year.

Automated coding refers to the process whereby instead of a clerk assigning a numeric code to a write-in answer, the respondent's answer is data captured and through use of reference files and matching algorithms the computer assigns a code. This differs from computer assisted coding where, while the computer may suggest a list of possible codes, the final decision is made by a human. Statistics Canada's monthly Labour Force Survey has had an operational autocoding system for industry and occupation for several years.

Recent research work on an autocoding system for the Census of Population began with a research study resulting in a report in March 1983. Statistics Canada personnel visited the USBC in July 1983 to observe the Hellerman I & O coding system. During the fall and winter of 1983-84, a prototype system for

the coding of industry was developed and tested. The exact question which was coded read as follows:

41.(a) For whom did you work?

_____

Name of firm, government agency, etc.

_____

Department, branch, division, section or plant

(b) What kind of business, industry or service was this?

_____

In addition, Question 42 asked for address where the person worked.

The prototype system consisted of two subsystems:

(i) The Business Register (BR) subsystem, which used a machine readable list of establishments with industry codes already assigned.

(ii) The Standard Industrial Classification (SIC) subsystem.

The process flow diagram (Figure 1) illustrates the various steps in the attempt to assign a code.

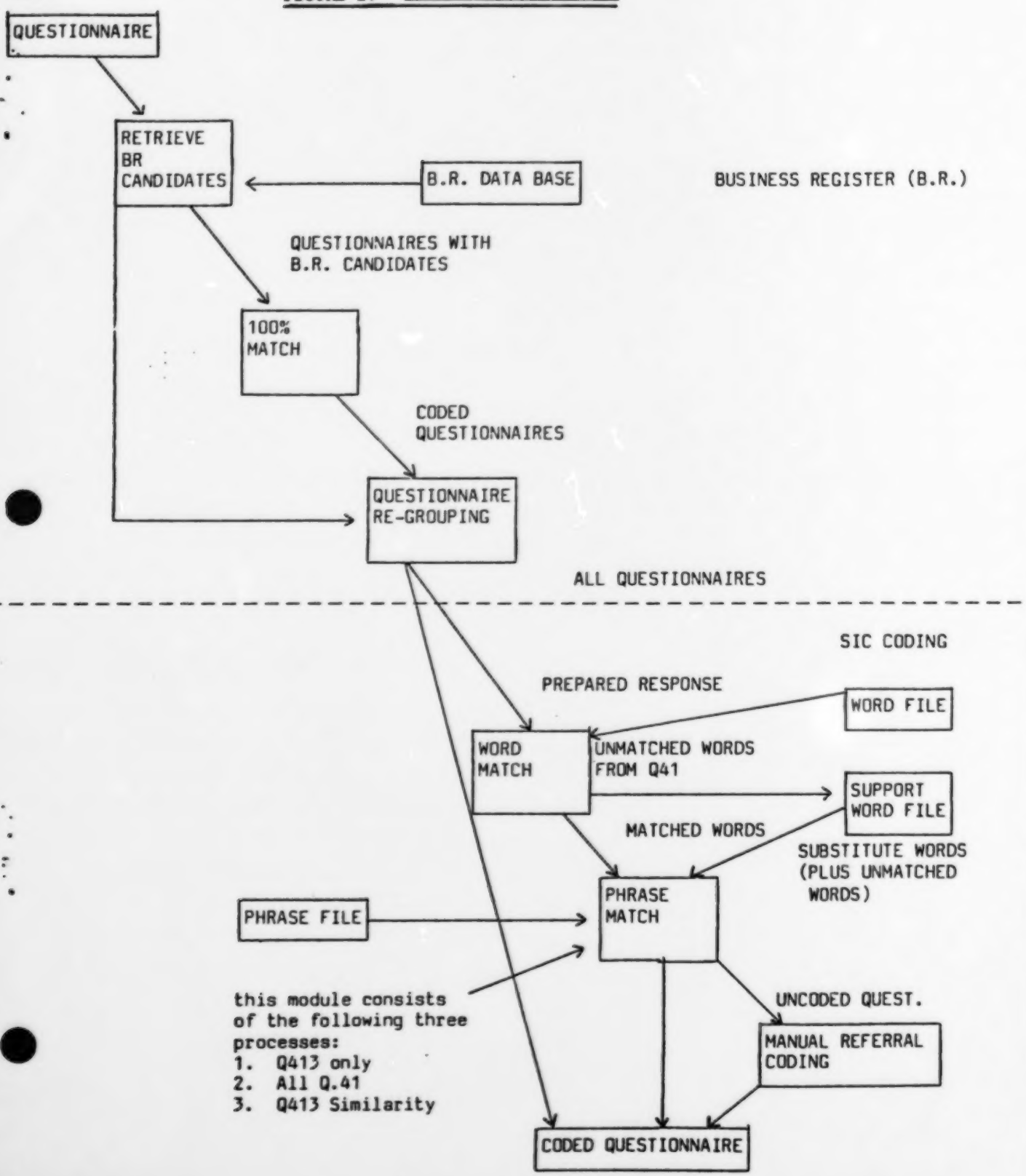
BR Subsystem

In this subsystem, similar records were selected from the BR database for each write-in, the BR records were compared to the company name and address in a 100% match, and if an exact match was found the corresponding SIC code was assigned. If no match was found, the questionnaires were passed on to the SIC subsystem.

SIC Subsystem

The first step in this subsystem was to prepare the write-in responses for the Word Match module by standardizing all words (e.g., removing suffixes and double letters, changing plural forms to singular) and by removing all punctuation and trivial words (e.g., the, and, or). The Word Match module

FIGURE 1: PROCESS FLOW DIAGRAM



then involved a search of the word file to find an exact match for each word in the Question 41 response. For all words not matched, the system then searched a support word file for a substitute word. This file contained a spelling corrector, list of synonyms and abbreviations and a list of foreign words. For each word found in these searches, the heuristic weight of the word was assigned. The heuristic weight is basically a measure of the uniqueness of the word and is a function of the entropy of the word; the more unique the word, the higher the weight.

All matched words from the Word Match then went through the Phrase Match module. The words were first arranged in decreasing order by weight. Beginning with the word with the highest weight, the system selected all phrases containing this word. If all the words in the respondents answer matched all the words in one of the phrases, the corresponding code was assigned. Otherwise each phrase was scored and the score compared to set criteria. If one of the phrases met the criteria, the code was assigned, otherwise the system tried the whole process again using the word with the second highest weight.

The Phrase Match actually consisted of three different processes:

(i) a match on just the third part (kind of business) of Question 41

(ii) a match on all three parts of Question 41

(iii) a "similarity match" using the kind of business part of Question 41.

For process (iii), the system took the unmatched words from the kind of business part of the Word Match module and searched the word file for a "similar" word, using a bit-string comparison. These similar words and the matched words from the third part of Question 41 were used in a final Phrase Match attempt.

The results of the prototype test are shown in Table 4. Two test runs were made, which differed only in what was matched in the BR subsystem. In the first run, the company name and full address were used, while in the second run only the company name and province were used.

The match rates achieved by the prototype system are not at the level

where automated coding would be a viable alternative to manual coding. The error rates of the machine coding are also above that of a manual coding operation. A complete analysis of the reasons has not yet been completed, but the primary reason appears to be deficiencies in the reference files. With improvements in the reference files, refinement in matching algorithms, and the

Table 4: MATCH RATES AND ERROR RATES by REPLICATION AND MODULE

Module	Co. Name and Address			Co. Name and Province		
	No.	Match Rate (%)	Error Rate (%)	No.	Match Rate (%)	Error Rate (%)
BR Subsystem						
Records for coding	3811			3811		
SIC code found	475	12.5		1073	28.2	
SIC code in error	52		10.9	128		11.9
Kind of business module						
Records for coding	3572			2981		
SIC code found	533	14.9		435	14.6	
SIC code in error	98		18.4	74		17.0
All Q41 module						
Records for coding	3162			2662		
SIC code found	105	3.3		92	3.5	
SIC code in error	48		45.7	41		44.6
Total coding system						
Records for coding	4170			4170		
SIC code found	1113	26.7		1600	38.4	
SIC code in error	198		17.8	243		15.2

integration of the occupation variable, it should be possible to increase these rates substantially.

7. CONCLUSION

There has been space in this paper to describe only a few issues that we will be addressing for 1986. Among those not mentioned are those such as the use of telephone follow-up for complete nonresponse and possible applications

of subscriber lists from the telephone companies, cross-edit and imputation between 100% and sample variables, and revisions to the weighting system. Despite the fact that 1986 is considered to be a "no change" census, there still is lots of work to do. Or perhaps it is simply the old saying run backwards: "The more things stay the same, the more they change."

REFERENCES

[1] Census Operations Division, Report on the Feasibility of Integrating the Disability Survey into the Census Process, Census Operations Division, Statistics Canada, September 1984.

[2] Census Survey Methods Subdivision, A Presentation on the Methodology of the 1981 Census, Census and Household Survey Methods Division, Statistics Canada, June 1982.

[3] Cochran, W.G., Sampling Techniques, 3rd ed., John Wiley & Sons, New York, 1977.

[4] Fellegi, I., and Holt, D. A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association 71, pp. 17-35, 1976.

[5] Hollins, D., and Dibbs, R., Investigation of Nonresponse and Data Quality in Indian Reserves EA's, 1981 Census of Population and Housing, Census and Household Survey Methods Division, Statistics Canada, August 1984.

[6] Statistics Canada, Discussion Paper - Plans for the 1986 Census, Office of the Census Manager, Statistics Canada, August 1983 (revised November 1983).

[7] Statistics Canada, Automatic Coding Research and Development, Industry Prototype - Analysis Report, Census Operations Division, Statistics Canada, March 1984.

ACKNOWLEDGEMENTS

The author would like to thank Dave Hollins and Alex van Baaren for valuable assistance in the preparation of this paper.